Independent and Automatic Evaluation of Speaker-Independent Acoustic-to-Articulatory Reconstruction

Maud Parrot, Juliette Millet, Ewan Dunbar

Independent and Automatic Evaluation of Speaker-Independent Acoustic-to-Articulatory Reconstruction

Maud Parrot[1], *Juliette Millet*[1,2,3], *Ewan Dunbar*[1,2]

[1] CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
[2] Université de Paris, LLF, CNRS, Paris, France
[3]CRI, Département Frontières du Vivant et de l'Apprendre, IIFR, Université de Paris
juliette.millet@cri-paris.org

Abstract

Reconstruction of articulatory trajectories from the acoustic speech signal has been proposed for improving speech recognition and text-to-speech synthesis. However, to be useful in these settings, articulatory reconstruction must be speaker-independent. Furthermore, as most research focuses on single, small data sets with few speakers, robust articulatory reconstruction could profit from combining data sets. Standard evaluation measures such as root mean squared error and Pearson correlation are inappropriate for evaluating the speaker-independence of models or the usefulness of combining data sets. We present a new evaluation for articulatory reconstruction which is independent of the articulatory data set used for training: the *phone discrimination ABX task*. We use the ABX measure to evaluate a bi-LSTM based model trained on three data sets (14 speakers), and show that it gives information complementary to standard measures, enabling us to evaluate the effects of data set merging, as well as the speaker independence of the model.

Index Terms: articulatory inversion, speech representation, machine learning, bi-LSTM

1. Introduction

Acoustic-to-articulatory inversion is the problem of finding a mapping from acoustic features to a set of articulatory measures (see [1, 2, 3, 4, 5, 6] for recent models: see [7] for a review). Reconstructed articulatory trajectories have been shown to improve text-to-speech synthesis [8, 7], speech accent conversion [9], and automatic speech recognition [10, 4], in particular for dysarthric speech [11]; they can also be used in the automatic detection of clinical conditions which have an impact on speech production, such as Parkinson's [12]. *Speaker independent* reconstruction, which abstracts away from speaker-specific articulatory idiosyncrasies, is essential for most of these applications. As such, a method for evaluating speaker-independent articulatory reconstruction is needed.

The two principal metrics for evaluating articulatory reconstruction are the root mean square error (RMSE) and the Pearson correlation coefficient (PCC) between the reference and the predicted articulatory trajectories. However, the goal of a speaker-independent model is not precise prediction of reference measures. First, the shapes of speakers' vocal tracts vary in ways that cannot be compensated by simple normalization. Second, recording conditions (coil placement in electromagnetic articulography: EMA), vary across and within recordings. Third, acoustic-to-articulatory inversion is a one-to-many problem: multiple articulatory trajectories can produce the same sound [13]. Speaker-independent articulatory models should not be penalized for reconstructing trajectories which are possible but different from the reference, or abstract (for example, average) trajectories. Finally, reference articulatory data is needed to calculate these measures, which is costly to obtain.

We propose a standardized evaluation based on an *ABX phone discrimination test* [14] of trajectories reconstructed for an acoustic-only corpus. The evaluation has the advantage of being *independent* of the training data set and of the true reference articulatory trajectories, much like the independent evaluation of [15], which uses human listeners to evaluate speech re-synthesis on the basis of the reconstructed trajectories. Our evaluation, however, is completely *automatic*, and does not need any human test.[1]

We train a bi-LSTM model closely resembling that of [1] on three data sets (MOCHA–TIMIT, EMA–IEEE, USC–TIMIT), and apply the ABX phone discrimination evaluation on different training set, validation set, and test set. We compare ABX phone discrimination scores to standard metrics, and show that it can complete articulatory inversion models' evaluation when speaker independant reconstructions are wanted.

2. Method

2.1. Model architecture

To demonstrate our evaluation, we use a bidirectional recurrent neural network architecture similar to that of [1], but with the addition of a convolutional layer that acts as a low pass filter after the readout layer. The network has two feed-forward layers of 300 units each that act as feature extractors, two bidirectional layers of 300 units each, a convolutional layer as a low pass filter,[2] and, finally, a feed-forward layer with as many units as the number of trajectories predicted (see Figure 1, and details on the outputs in 3.2 and 3.3 below).

2.2. Loss function

The usual loss function for articulatory inversion is the root mean square error (RMSE), which minimizes the L2 distance between the real and predicted trajectory. The Pearson correlation coefficient (PCC) is also typically examined as an evaluation metric for articulatory inversion models, in addition to the RMSE values, as it ignores any systematic differences between speakers that can be described linearly. The PCC measures the

[1]All code for pre-processing the data sets and for training and testing the model is fully available at `https://github.com/bootphon/articulatory_inversion.git`

[2]As [1] shows, this is not strictly necessary in order to obtain smooth trajectories. See section 2.3 for discussion.

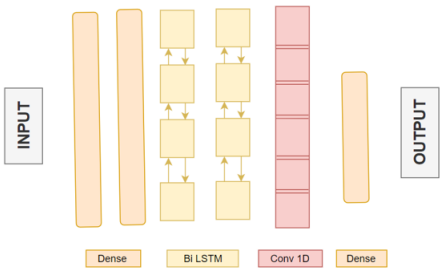

Figure 1: *Neural network architecture used in this paper.*

degree of linear relationship between two variables X and Y.

$$PCC(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

We experiment with including the PCC in the loss function, y the reconstructed trajectory y^* the reference:

$$\mathcal{L}(y,y^*) = RMSE(y,y^*) - \beta \cdot PCC(y,y^*) \tag{2}$$

We use $\beta = 1000$ (in order to account for the differences in scale between the RMSE and PCC values).[3]

2.3. Convolutional layer as low pass filter

Articulatory gestures are not only continuous but also smooth. We impose smoothing on the predicted output trajectories by integrating a convolutional layer that acts as a low pass filter at the output of our neural network. This avoids unnecessary backpropagation of error due to non-smooth predicted trajectories. We used the following filter of order 5, with a Hanning window to restrict the support, where N is the size of the Hanning window and $f_t, \forall n \in [0, N-1]$:

$$w(n) \propto \quad \left(1 - \cos(2\pi \frac{n}{N-1})\right) \operatorname{sinc}\left(2\pi f_t (n - \frac{N-1}{2})\right) \tag{3}$$

2.4. ABX phone discrimination evaluation

An *ABX phone discrimination test* of a representation of speech consists of extracting the representations of triplets of stimuli (A, B, and X), and computing the distance $d(\mathrm{A},\mathrm{X})$, between A and X, and $d(\mathrm{B},\mathrm{X})$, between B and X. X is of the same phonetic category as either A or B. Taking A to be the correct answer, we compute $\delta = d(\mathrm{B},\mathrm{X}) - d(\mathrm{A},\mathrm{X})$. If $\delta > 0$, the model has chosen A; if $\delta < 0$, it has chosen B. As in previous work evaluating acoustic models with this method [16, 17], the percent correct for all pairs of categories are combined into a global ABX discriminability score.

In our case, the stimuli are triphones (like [seɪk] or [zfɑ], see 4.4 below for more details). The representations we extract are articulatory trajectories reconstructed from these stimuli. These representations are $N_c \times a$ matrices for stimulus c, with a the number of articulator positions our model reconstructs, and N_c the number of time frames the stimuli is. Not all stimuli are of the same length (stimuli are time sequences, and so N_c can change from one to the other), and so computing the distance between two stimuli representation is not straightforward. We follow previous literature and use dynamic time warping (DTW) to align sequences of differing length [18]. Dynamic time warping takes two sequences C and D as input, as well as a function γ for comparing pairs of sequence elements. It aligns C and D by matching the elements of one to the other so as to minimize the sum of $\gamma(c,d)$ for all matched elements (c,d). Each element of C must be matched with at least one element of D, and alignments must respect temporal order. Once the matching is decided, the distances between stimuli $C = c_1, c_2, ...c_{N_c}$ and $D = d_1, d_2, ...d_{N_d}$ is:

$$d(C,D) = \frac{\sum_{c_i, d_j \text{ are matched}} \gamma(c_i, d_j)}{max(N_c, N_d)} \tag{4}$$

We use the following cosine distance as γ:

$$\gamma_{cos}(\mathbf{x},\mathbf{y}) = \frac{1}{\pi} \arccos\left(\frac{\sum_{i=1}^{A} x_i y_i}{\sqrt{\sum_{i=1}^{A} x_i^2}\sqrt{\sum_{i=1}^{A} y_i^2}}\right) \tag{5}$$

where $\mathbf{x}$ and $\mathbf{y}$ are individual time frames containing a reconstructed articulatory measures.

Articulatory reconstruction should predict different representations for different phones. The labels we use are phonemic, but controlling for immediate left and right single-segment context serves to (imperfectly) limit variability due to allophony or coarticulation. The ABX phone discrimination error (the percentage of cases for which δ picks out the incorrect stimulus) should be low for good articulatory inversion. Speaker-independent articulatory reconstruction should have low phone discrimination error even when reconstruction is done on novel test speakers. Furthermore, speaker-independent articulatory reconstruction should give representations which are stable across speakers, which means that scores should also be low when X is uttered by a different speaker than A and B.

3. Articulatory data sets

3.1. Description

All the data used are freely available. MOCHA–TIMIT[4] is a database that contains EMA and acoustic data for 460 utterances (20 min) read by two English speakers [19]; USC–TIMIT[5] [20] provides EMA data for four speakers on the MOCHA–TIMIT sentences (15 min); and EMA–IEEE[6] [21] contains eight speakers reading 720 sentences, once each at a normal rate, and once at a fast speech rate (47 min per speaker).[7] The combined duration is 461 min.

3.2. Articulatory trajectories

We use measures in the sagittal plane (x: back to front of head; y: chin to forehead). The two-dimensional articulatory points available in our data sets are: tongue body (TB), tongue tip (TT), tongue dorsum (TD), upper lip (UL), lower lip (LL), lower incisor (LI), and velum (V). Not all measures are available for all speakers. The velum trajectory is only available in MOCHA–TIMIT, and we exclude specific articulators for certain speakers where the standard deviation is less than 0.5 mm

[3] We ran experiments varying the weights of the RMSE and the PCC, but found it had no systematic effect on the results.

[4] http://data.cstr.ed.ac.uk/mocha

[5] https://sail.usc.edu/span/usc-timit/

[6] https://yale.app.box.com/s/cfn8hj2puveo65fq54rp1ml2mk7moj3h/folder/30415804819

[7] We also conducted held-out tests on the single-speaker MNGU0 database, which we do not report due to preprocessing issues.

and visual verification suggested strongly that the measure was wrong. In training conditions combining speakers within corpora, we use the common articulators.

3.3. Vocal tract parameters

As in previous works [22, 23], we add vocal tract variables, using slightly different formulas from [22]. We calculate two tract variables from the position of the lips, the vertical lip aperture (VLA) and the horizontal lip protrusion (HPRO):

$$\text{VLA} = \text{UL}_y - \text{LL}_y \quad (6) \qquad \text{HPRO} = \frac{\text{UL}_x + \text{LL}_x}{2} \quad (7)$$

We also add the tongue tip constriction (TTC: the cosine of the angle of the tongue tip off the horizontal axis) and the tongue body constriction (TBC: the cosine of the angle of the tongue body off the horizontal axis).

$$\text{TTC} = \frac{\text{TT}_x}{\sqrt{\text{TT}_x^2 + \text{TT}_y^2}} \quad (8) \qquad \text{TBC} = \frac{\text{TB}_x}{\sqrt{\text{TB}_x^2 + \text{TB}_y^2}} \quad (9)$$

4. Experiments

We compare standard reconstruction scores against ABX scores as evaluations of speaker-independence, both within and across corpora. Within corpus, we compare models trained in a **speaker-specific** setting, on a single speaker, with models trained in a **speaker-independent** setting, training on multiple speakers. We randomly hold out data in the speaker-specific setting for validation and test (for calculating the reconstruction scores: 70% train, 10% validation, 20% test). In the speaker-independent setting, we validate on a subset of the speakers, test on one speaker, and train on the rest. The speaker-specific model gives an expected upper bound on reconstruction. We expect speaker-independent models to give poorer reconstruction, but seek to use the ABX score to assess whether this degradation is due to failure to reconstruct linguistically relevant articulatory information, or failure to reconstruct speaker-specific detail.

To test the effect of merging corpora, we compare a model trained in a **multi-corpus** setting (with speakers EMA–IEEE: M01, MOCHA–TIMIT: FSEW0, and USC–TIMIT: M1 held out for validation and test) against a **single-corpus** setting, training on EMA–IEEE, which contains the most complete set of articulators (speaker M01 still held out for test). Here, rather than training only on articulators common to all speakers, we learn to reconstruct all trajectories by ignoring error on missing articulators for backpropagation.

4.1. Model parameters

We use the Adam optimizer with early stopping on the validation set (learning rate 0.001, batch size 10, patience 5). The weights of the low pass filter are fixed according to (3) with $N = 50$ to give a transition band of 0.08. The convolution has one channel, stride of 1, and padding such that the output has the same size as the input. The cutoff frequency f_c is 10Hz.

4.2. Data preprocessing

We use as input the thirteen first MFCCs + Δ + $\Delta\Delta$ with window size of 25ms and stride of 10ms. We add 10 context windows: the five previous and five following frames, as in [24]. We remove silences based on the transcription file (when available). We normalize the MFCCs per speaker, removing the mean and dividing by the standard deviation.

We pre-smooth the articulatory trajectories, applying a low pass filter with a cutoff frequency of 10Hz for all the data sets except for EMA–IEEE, for which we use 20Hz. We remove leading and trailing silences ends using the transcription when available. We reduce EMA sampling rates to 100Hz to have a single articulatory frame per MFCC frame. Since EMA coils move gradually during recording [25], we normalize each articulatory measure by subtracting the mean over the 60 previous and following recordings of the same speaker and divide by the speaker-specific standard deviation.

4.3. Reconstruction scores

RMSE (mm) and RMSE computed on normalized trajectories are computed on every feature except TTC and TBC, and PCC is computed on every feature available.

4.4. ABX scores

The ABX test is performed on the one-second English (speech-only) test data set from the Zero Resource Speech Challenge 2017 [17], consisting of data from 24 speakers taken from the LibriVox audio book collection, labelled using the 39 CMU-DICT phonemes plus `ax` for [ə]. Stimuli are triphones differing in the central phone (*beg–bag*, *api–ati*, etc). *Within-speaker* triplets contain three triphones from a single speaker (e.g., $A = \text{beg}_{T1}$, $B = \text{bag}_{T1}$, $X = \text{bag}'_{T1}$). In *across-speaker* triplets, A and B come from the same speaker, and X to another. $A = \text{beg}_{T1}$, $B = \text{bag}_{T1}$, $X = \text{bag}_{T2}$. The scores for a given contrast are first averaged across all (pairs of) speakers for which triplets can be constructed, before averaging over all contexts, and over all pairs of central phones and being converted to an error rate by subtracting from 1. We exclude contrasts with less than three contexts and for which critical articulators were missing from the data.[8]

5. Results

The model attains reconstruction scores on speaker-specific training for FSEW0 which are comparable to existing results (RMSE-mm: 1.43, RMSE-norm: 0.55, PCC: 0.77).

5.1. Speaker-independence within corpus

Table 1 compares the average scores across speakers in the speaker-specific setting, versus the average over all one-speaker-held-out training configurations in the speaker-independent setting. As expected, the speaker-independent condition shows degradation in the three reconstruction scores, compared to the speaker-specific condition.

The ABX phone discrimination scores, calculated on an external speech corpus, provide a different picture. Unlike for the reconstruction scores, the ABX scores show only very small differences between the two training conditions for MOCHA–TIMIT and USC–TIMIT. Although we lose information about detailed articulatory tracks in the speaker-independent condition, this information is evidently not relevant to coding phone contrasts. In the EMA–IEEE corpus, where the difference in reconstruction scores between speaker-specific and speaker-independent conditions is smaller, the ABX scores show a

[8] For example, oral–nasal contrasts such as [ana]–[ada]–[ana], which depend necessarily on the position of the velum: a complete list is provided at `https://github.com/bootphon/articulatory_inversion`

Table 1: *Comparison between speaker-specific (**Sp**) and speaker-independent (**Ind**) settings. R: RMSE, the smaller the better, minimum value 0, no maximum value. Rn: normalized RMSE, the smaller the better, minimum value 0, no maximum value, but this value is comparable from one dataset to the other. PCC: Pearson Correlation Coefficient, between - 1 and 1, the bigger the better. A-w: within-speaker ABX percentage error, A-a: across-speaker ABX percentage error, between 0 and 100, the smaller the better. Scores are averages across training subsets (see section 4).*

	MOCHA–TIMIT					**USC–TIMIT**					**EMA–IEEE**				
	R	Rn	PCC	A-w	A-a	R	Rn	PCC	A-w	A-a	R	Rn	PCC	A-w	A-a
Sp	**1.380**	**0.557**	**0.759**	**23.9**	32.2	**1.478**	**0.608**	**0.747**	24.5	**33.8**	**1.557**	**0.501**	**0.840**	22.1	30.5
Ind	2.184	0.851	0.417	24.6	**32.0**	2.310	0.917	0.199	**24.3**	33.9	2.198	0.688	0.672	**18.4**	**24.8**

Table 2: *Effects of training one or multiple corpora. The different measures are detailed in Table 1. The speakers indicated on top are the speakers on which the measures are computed for each type of model. The last column indicates ABX percentage error within (w) and across (a) speaker on the dataset described in Section 4.4*

	M01 (EMA–IEEE)			**M1 (USC–TIMIT)**			**FSEW0 (MOCHA)**			**ABX**	
	R	Rn	PCC	R	Rn	PCC	R	Rn	PCC	w	a
Single-corpus	**1.79**	0.66	**0.72**	2.05	1.13	0.02	2.72	1.11	0.08	**18.9**	**25.0**
Multi-corpus	1.80	0.66	0.71	**1.89**	**1.04**	**0.14**	**2.61**	**0.98**	**0.22**	19.7	26.7

marked improvement in the speaker-independent condition. The speaker-independent model trained on this corpus not only does not lose information, but in fact better reconstructs the articulatory invariants necessary to code phone contrasts. This fact is not captured by looking only at the reconstruction scores, and can only be seen in the ABX phone discrimination measure.

5.2. Merging corpora

Results comparing multi-corpus with single-corpus training are shown in Table 2. Each row represents a single training condition (trained either on EMA–IEEE, for the **single-corpus** condition, or on all corpora, for the **multi-corpus** condition, in which the USC–TIMIT and MOCHA corpus are added to the training set. Each of the first three groups of columns gives reconstruction scores on each of the three held-out test speakers. With the addition of the two corpora in the **multi-corpus** condition, the reconstruction measures improve for the unseen test speakers drawn from the two newly added corpora. The fact that we observe appreciable improvement exclusively in the novel corpora suggests that the improvements in reconstruction are not simply due to the addition of more data, but are in part due to improved modelling of acoustic channel or coil placement properties specific to these corpora. However, the external ABX scores in the **multi-corpus** condition are similar to those in the **single-corpus** condition—in fact, slightly worse—suggesting that adding speakers from additional corpora to training is not beneficial to reconstruction of articulatory invariants, and that reconstruction of linguistically irrelevant properties of the data may be the *only* improvement the additional data provides. Note that improved reconstruction is unlikely to be due to improved modelling of speaker-level articulatory idiosyncrasies, since none of the test speakers appear in either training condition. Thus, while the reconstruction scores might initially suggest a benefit to the **multi-corpus** training, the external phone discrimination score again reveals a more complete picture, casting doubt on any such benefit for the purposes of speaker-independent modelling.

6. Conclusion

We have proposed an ABX phone discrimination measure for the evaluation of speaker-independent acoustic-to-articulatory models. The measure is independent of the articulatory trajectories, and thus does not penalize models for failing to capture speaker-specific articulatory details; it comes from a single external corpus, avoiding the inherent instability of held-out measures; and it is automatic, unlike speech-synthesis based evaluations. Our ABX score only assesses the presence of information needed to contrast the phones labelled in the corpus used (40 English phoneme labels), but they can be replaced by finer-grained allophonic labels, if desired. One caveat of phone discriminability is that it will vary as a function of the set of articulatory dimensions reconstructed, not only of how well they are reconstructed. So this measure can only be used to compare models that reconstruct the same articulators. Nevertheless, we have shown that it can give important information complementary to traditional reconstruction scores, indicative of the degree to which improvements or declines in reconstruction are due to failure to reconstruct speaker-specific properties.

7. Acknowledgements

This research was supported by the École Doctorale Frontières du Vivant (FdV) – Programme Bettencourt, by Facebook AI Research, and by grants ANR-17-CE28-0009 (GEOMPHON), ANR-11-IDFI-023 (IIFR), ANR-18-IDEX-001 (UdP), ANR-10-LABX-0083 (EFL).

8. References

[1] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4450–4454.

[2] L. Yu, J. Yu, and Q. Ling, "Synthesizing 3d acoustic-articulatory mapping trajectories: Predicting articulatory movements by long-term recurrent convolutional neural network," in *2018 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2018, pp. 1–4.

[3] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1274–1277.

[4] V. Mitra, G. Sivaraman, C. Bartels, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franco, "Joint modeling of articulatory

and acoustic spaces for continuous speech recognition tasks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5205–5209.

[5] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion," *Proc. Interspeech 2019*, pp. 859–863, 2019.

[6] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," *arXiv preprint arXiv:1904.06083*, 2019.

[7] K. Richmond, Z. Ling, and J. Yamagishi, "The use of articulatory movement data in speech synthesis applications: An overview—application of articulatory movements using machine learning algorithms—," *Acoustical Science and Technology*, vol. 36, no. 6, pp. 467–477, 2015.

[8] B. Cao, M. J. Kim, J. P. van Santen, T. Mau, and J. Wang, "Integrating articulatory information in deep learning-based text-to-speech synthesis." in *INTERSPEECH*, 2017, pp. 254–258.

[9] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7694–7698.

[10] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," 2000.

[11] E. Yılmaz, V. Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent asr of dysarthric speech," *Computer Speech & Language*, vol. 58, pp. 319–334, 2019.

[12] S. Hahm and J. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[13] P. Delattre and D. C. Freeman, "A dialect study of american r's by X-ray motion picture," *Linguistics*, vol. 6, no. 44, pp. 29–68, 1968.

[14] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFCCs/PLP pipeline," 2013.

[15] K. Richmond, Z.-H. Ling, J. Yamagishi, and B. Uria, "On the evaluation of inversion mapping performance in the acoustic domain." in *INTERSPEECH*. Citeseer, 2013, pp. 1012–1016.

[16] M. Versteegh, R. Thiollière, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015," in *INTERSPEECH-16*, 2015.

[17] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The Zero Resource Speech Challenge 2017," in *ASRU 2017*. IEEE, 2017, pp. 323–330.

[18] P. Senin, "Dynamic time warping algorithm review," 2008, ms., Department of Information and Computer Sciences, University of Hawaii. [Online]. Available: http://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf

[19] A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," 2000.

[20] S. Narayanan, A. Toutios, V. Ramanarayanan, A. C. Lammert, J. Kim, S. Lee, K. S. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. K. Ghosh, A. Katsamanis, and M. I. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)." *The Journal of the Acoustical Society of America*, vol. 136 3, p. 1307, 2014.

[21] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017. [Online]. Available: https://doi.org/10.1121/1.4987629

[22] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, pp. 316–329, 07 2019.

[23] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion." in *INTERSPEECH*, 2016, pp. 455–459.

[24] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Proceedings of Interspeech*. Curran Associates, 2012, pp. 866–870.

[25] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, University of Edinburgh, 2002.